Sentiment Analysis

A Probabilistic Approach

February 1, 2012

Supervised by I. Langbroek (Blauw Research)
Mentored by dr. M. van Someren (Universiteit van Amsterdam)

S. A. Gieske	S. Laan	C. R. Verschoor	D. S. ten Velthuis	A. J. Wiggers
6167667	6036031	10017321	0577642	6036163

Artificial Intelligence
Faculty of Science
Universiteit van Amsterdam

Contents

1 Introduction

With the growth of Social Media, such as social networks and blogs consumers gained a place to review, rate, recommend products on-line. Making on-line opinion important for companies who want to market their products, manage their reputations and identify new opportunities. Being able to automate the process of finding relevant content, filtering out noise and categorize the messages will speed up the process compared of doing it by hand. Research has shown that sentiment analysis is showing promising results to automate this process.

Sentiment analysis is the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials. Which makes the data from social media ready for further analysis. Using sentinemt analysis we will group twitter messages into positive, negative and neutral messages

2 Problem Specification

Our dataset was provided by our client, a excel file excisting of around 10.000 social messages (blogs,facebook post and twitter messages), all graded by hand into 5 categories: +2 very positive, +1 positve, 0 neutral, -1 negative, -2 very negative. The goal of our project is to build a learning algotrm that can classify these messages by itself.

2.1 Data Cleaning

A persons tweet can only contain 140 characters, so to still be able to express yourself users mostly use slang and emoticons. Which increases the value of cleaning the messages first. For example punctuations such as: dot's, comma's and list's dont add anything to the sentiment of a messages so we deleted all useless information. Exclamation and question marks do give value to the sentiment of a sentence because, they can tell something about the emotion of the sentence so we left those in our corpus.

Making use of a Dutch stemmer found in the NLTK database (heb ik nog niet over gehad uitleg nodig), we are able to reduce inflected words to their stem, base or root form. Most words have the same meaning, for example "FUCK and Fucking" are related to each other and by stemming the 2 words to their base form "fuck" we decrease the amount of different words in our corpus and gain more same words (moet beter verwoord worden).

(iets vertellen over tokenizen? dat we een lijst maken van onze woorden?)

2.2 Data Reduction

Because, of the difference between blogs, facebook post and twitter feeds, we decided to only focus on the twitter feeds. Blogs, have larger text than the other two and have a good change that different topics and more opinons are discussed than with twitter. Increasing the need for a different aproach for blogs than twitter. Facebook messages

look more like twitter but, beacuse of the restriction on the size of twitter messages there is need for a short writting style that is not needed in the other 2 messages

2.3 Machine Learning

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

3 Theory

3.1 Relevant Literature

3.2 Machine Learning

3.2.1 Perceptron

3.2.2 Artificial Neural Network

3.2.3 Method 3

3.3 Problistics Terms

3.3.1 Accuracy

3.3.2 Precision

3.3.3 Recall

3.3.4 F-Measure

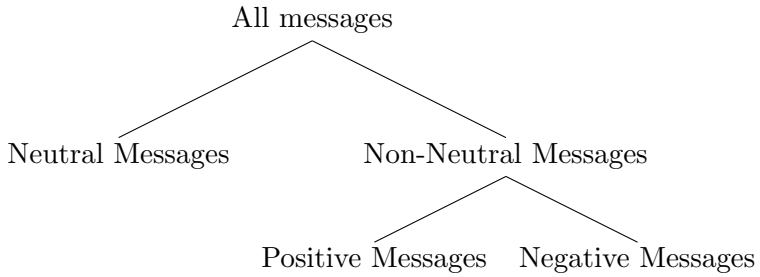
4 Usage and User Guide

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5 Implementation and Results

5.1 Global Approach

First we'll explain our general approach. The idea is to use a binary classifier, i.e. either true or false, multiple times. This way we can first decide whether a message is neutral and then we can decide between the non-neutral messages whether it is a positive or negative message. Maybe a tree structure shows the used approach more clearly:



As you can see, with this approach we need only to distinguish between two classes at a time. This is very convenient, because there are a lot of techniques that can discriminate between two classes, whereas multi-class-classifiers are less common.

5.2 Binary Classification

In this section, we'll describe the method that we use to classify a message. The approach consists of a few steps:

1. The counting of word frequencies
2. Calculation of word probabilities
3. Calculation of message probabilities
4. Finding a good threshold

First we count the frequency of each word, i.e. we count how many times it occurs in the training set. We also keep track of how many times this word was encountered in either class. For example, if we want to decide whether messages are neutral or not, we keep track of how many times each word has been seen in a neutral message. The number of times the word occurs in the opposite class, is equivalent to the total number of encounters, minus the encounters in the first class.

When the frequencies of all words[1] are found, we can calculate a probability for each word, which gives us an idea of how likely it is to be encountered in a certain class. The formula for this probability is the following:

[1] All words in the training set

$$P(word) = \frac{\sum word \in C_1}{\sum word \in C_1 \cup C_2} \tag{1}$$

So the probability that a word is in C_1, the first class, is the number of times it has been encountered in a sentence, which was tagged to be in C_1, divided by the total number of encounters.

Now that all words have probabilities assigned to them, or at least all words in the training set, we can calculate the probabilities of the sentences.

$$P(s) = \frac{1}{n} \sum_{w \in s} P(w) \tag{2}$$

Where s is the sentence, n the number of words in the sentence and w a word in the sentence. As follows from the formula, we take the weighted sum of the word probabilities.

Now that every sentence has an associated probability, the real machine learning can begin. The goal is to find a threshold for the probabilities, i.e. when a sentence probability is higher than a certain amount, we classify it as C_1, otherwise, we classify it as belonging to C_2.

We tried several approaches to accomplish this task.

5.3 Algoritme 1

5.3.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique

quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4 Algoritme 2

5.4.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

6 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique

quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

7 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.